Characterising X-ray data sets with *Xtriage*

P.H. Zwart*[a], R.W. Grosse-Kunstleve[a], R.J. Read[b] & P.D. Adams[a] email: PHZwart@lbl.gov

a: Lawrence Berkeley National Laboratory, 1 Cyclotron Road, BLDG 64R0121 Berkeley, California 94720-8118, USA
b: Department of Haematology, University of Cambridge, Cambridge Institute for Medical Research, Cambridge, CB2 2XY, UK

Abstract

A set of comprehensive data quality checks is presented that allows the user to obtain an overview of the quality of the merged data. Various statistics (such as I/σ, completeness, Wilson-plot sanity and a comprehensive twinning analysis) are listed.

1. Introduction

With the emergence of structural genomics, more effort is being invested into developing methods that incorporate basic crystallographic knowledge to enhance decision making procedures (Adams *et al.*, 2004). The structure solution process can often be enhanced by having a basic understanding of certain characteristics of the X-ray data set under investigation. For instance, detecting the presence of anisotropic diffraction or twinning while a crystal is on the beam line, may allow the user to change the data collection strategy in order to obtain a better or a more complete data set. In post-collection analysis, the presence of (for instance) non-crystallographic translational symmetry might help the user to solve the structure more easily.

A number of checks have been implemented in the program *Xtriage*. *Xtriage* provides a concise overview of a wide variety of statistics that characterize an X-ray data set. Here a selected number of features of *Xtriage* are highlighted.

2. Measurability of anomalous signal

The presence of anomalous signal is tested for by determining the fraction of significant Bijvoet intensity differences. A Bijvoet intensity difference is qualified as significant if the following two conditions are met

$$\min\left[\frac{I_+}{\sigma_+}, \frac{I_-}{\sigma_-}\right] \geq 3 \qquad (1)$$

$$\frac{|I_+ - I_-|}{\sqrt{\sigma_+^2 + \sigma_-^2}} \geq 3 \qquad (2)$$

The fraction of Friedel mates that fulfill the above conditions is referred to as the measurability. Values of the measurability above 3% can be evidence for the presence of sufficient signal for substructure solution.

Figure 1 illustrates the resolution dependence of the measurability for a 4 wavelength MAD data set. All wavelength besides the `low energy remote` have enough anomalous signal to determine the substructure.

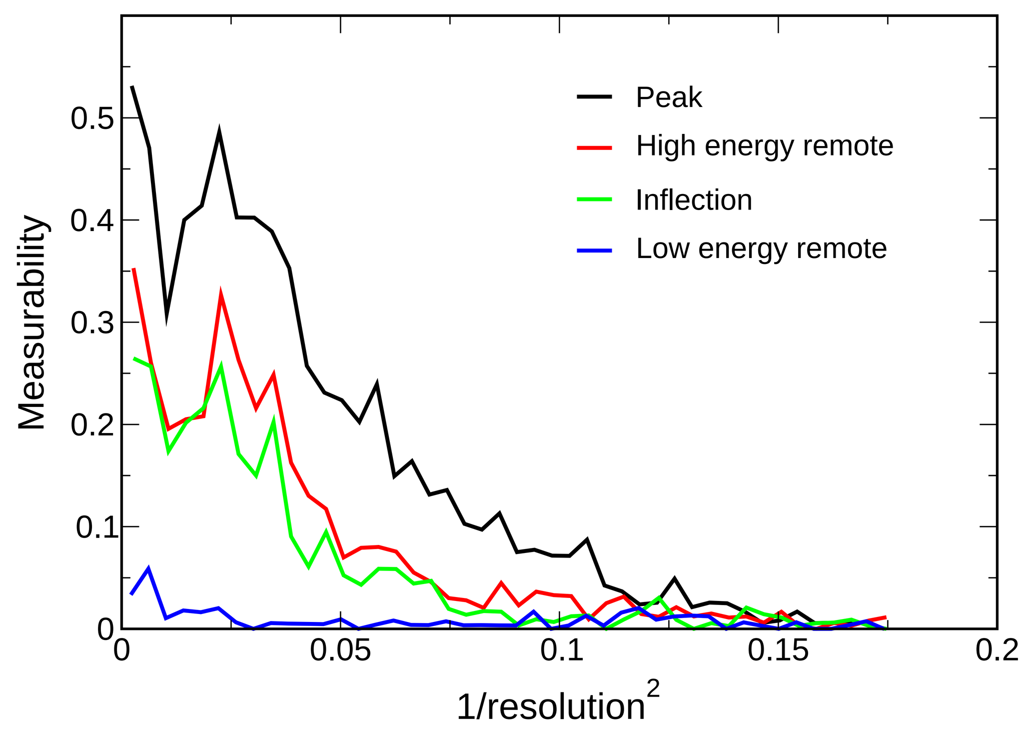

Figure 1. The measurability as a function of resolution for a 4 wavelength data set (PHENIX example nsf-d2-mad).

3. Wilson scaling

In order to reduce the impact of anisotropy of the Xray data on the intensity statistics, an anisotropic scale factor is determined, using likelihood methods as suggested by Popov & Bourenkov (2000). The resulting anisotropic tensor can be used to 'correct' the data. The effect of this correction on the intensity statistics is shown in Figure 1.

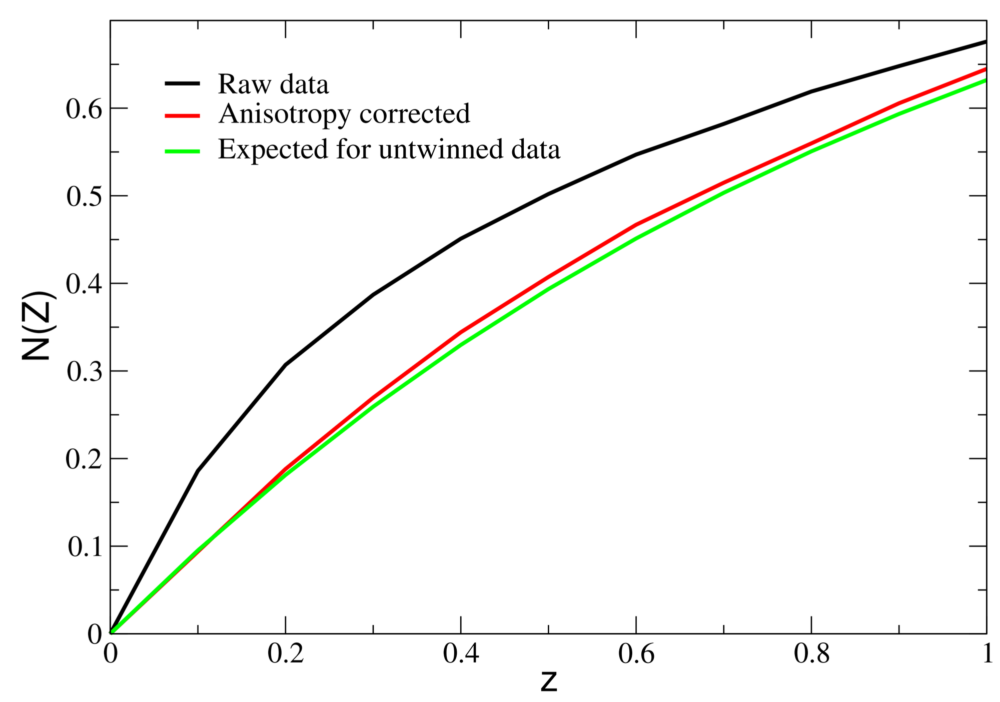

Figure 2. The effect of anisotropy correction on the cumulative intensity distribution. The estimated anisotropic $\mathrm{B}_{\mathrm{cart}}$ tensor was equal to (75,75,160,0,0,0)

References

Adams PD, Gopal K, Grosse-Kunstleve RW, Hung L-W, Ioerger TR, McCoy AJ, Moriarty NW, Pai RK, Read RJ, Romo TD, Sacchettini JC, Sauter NK, Storoni LC, Terwilliger TC (2004). *J. Synchrotron Rad.*, **11** 53–55.

Pannu, NS & Read, RJ (1996). *Acta Cryst.* **A52**, 659–668.

Popov, A & Bourenkov, G (2003). *Acta Cryst.* **D59**, 1145–1153.

Sauter NK, Grosse-Kunstleve RW & Adams PD.(2006). *J. Appl. Cryst.* **39**, 158–168.

Zwart PH, Grosse-Kunstleve RW, Adams PD (2005). *CCP4 newsletter* **42** 10.

4. Likelihood based estimation of twin fraction

When twin laws are available, the twin fraction of the data is typically estimated with a Britton plot, or via an H-test (e.g. Zwart *et al.*, 2005). A drawback of these methods is that they fail to take into account the effects of experimental errors, as well as a possible correlation between twin related intensities due to non-crystallographic symmetry. We have developed a likelihood based approach that aims to account for the influence of both factors upon estimation of a twin fraction.

Correlations between twin related intensities due to rotational NCS parallel to the twin axis, are treated by introducing non-zero covariances in the joint probability distribution of twin related structure factors. After integrating out phases, a subsequent change of variables from amplitudes to twinned intensities results in a distribution of a twin related intensity pair (J_1, J_2):

$$P(J_1, J_2) = \frac{1}{(\epsilon\sigma_p)^2(1-D_{\mathrm{ncs}}^2(1-2\alpha))} \exp\left[\frac{I_1+I_2}{\epsilon\sigma_p(1-D_{\mathrm{ncs}}^2)}\right] I_0\left[\frac{2D_{\mathrm{ncs}}(I_1I_2)^{1/2}}{\epsilon\sigma_p(1-D_{\mathrm{ncs}}^2)}\right] \qquad (3)$$

$$I_1 = ((1-\alpha)J_1 - \alpha J_2)/(1-2\alpha) \qquad (4)$$

$$I_2 = ((1-\alpha)J_2 - \alpha J_1)/(1-2\alpha) \qquad (5)$$

In the above expression σ_p denotes the mean intensity, ϵ a symmetry enhancement factor and α the twin fraction. D_{ncs} models the effect of a correlation between twin related structure factors. If the twin fraction α is equal to zero, the correlation between twin law related intensities is equal to D_{ncs}^2. Experimental errors can be dealt with in a similar way as carried out by Pannu & Read (1996).

The effects of non crystallographic symmetry on the estimation of the twin fraction via a Britton plot, is shown in Figure 2. Expression (3) was used to obtain theoretical Britton plot given a twin fraction of 25% and various values of D_{ncs}. From Figure 2 it is clear that the presence of a correlation between untwinned intensities results in an overestimation of the twin fraction by means of a Britton plot.

An example of a case of twinning with rotational NCS parallel to the twin law can be found with the Xray data of PDBID 1KU5. Standard techniques such as the Britton plot or the H-test estimate the twin fraction to be about 30%. The likelihood based technique estimate the twin fraction to be 21%. A correlation analyses in which untwinned model data is artifically twinned and compared to the observed data suggests a twin fraction of 24%.

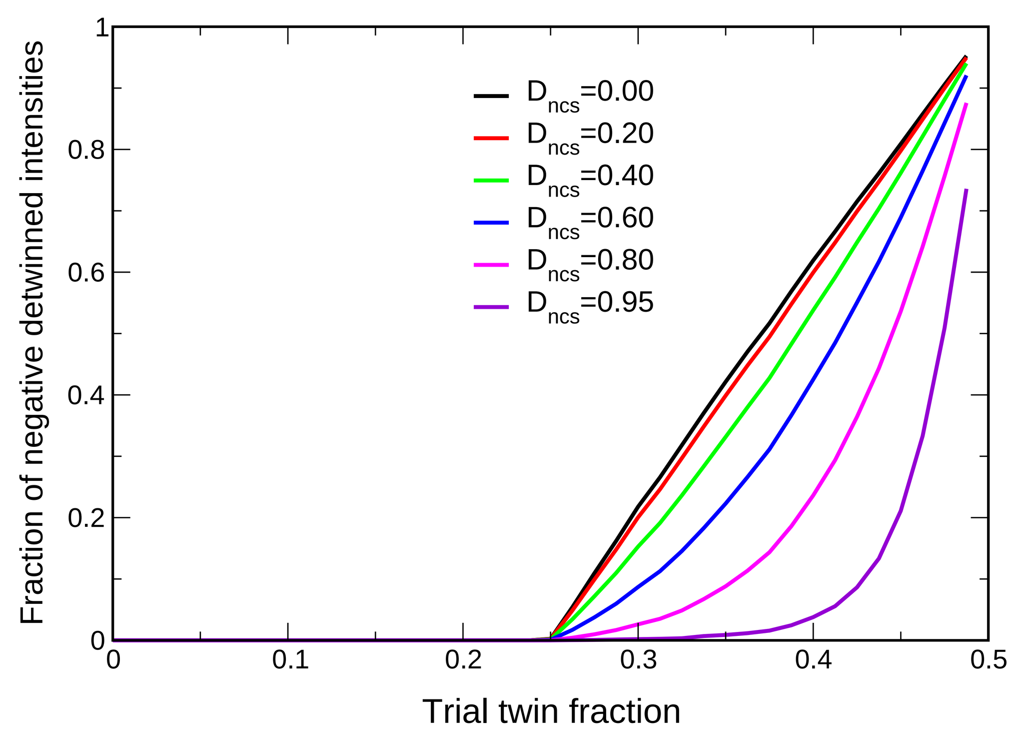

Figure 2. The effect of a correlation between twin related intensities due to NCS on a Britton plot. Shown is a theoretical Britton plot for a 25% twinned data set, without experimental errors, but with various values of D_{ncs}.

6. Missing rotational intensity symmetry

It is not uncommon that the assigned space group after data processing is too low. *Xtriage* tries to determine the point group of the intensities on the basis of an R-value analysis of intensities related by a specific symmetry operator, similar to the scheme discussed by Sauter *et al.* (2006). The list of possible point groups is presented as a graph that relates point groups via group-subgroup relations. For each possible point group, the set of symmetry operator of the lattice is partitioned in a set of *used* and *unused* operators. The *used* operators are elements of trial point group, whereas the *unused* operators are not. The most likely space group is identified by the a combination of low R-values for *used* symmetry operators and high R-values for *unused* symmetry operators.

As an example, a data set indexed and processed in `C2` with unit cell parameters `203 124 119 90 123 90` produces the point group graph shown in Figure 3. The R-value analysis reveals that the point group `R32:R` is most likely. The corresponding unit cell in the reference (hexagonal) setting is equal to `(123.548, 123.548, 284.709, 90, 90, 120)`.

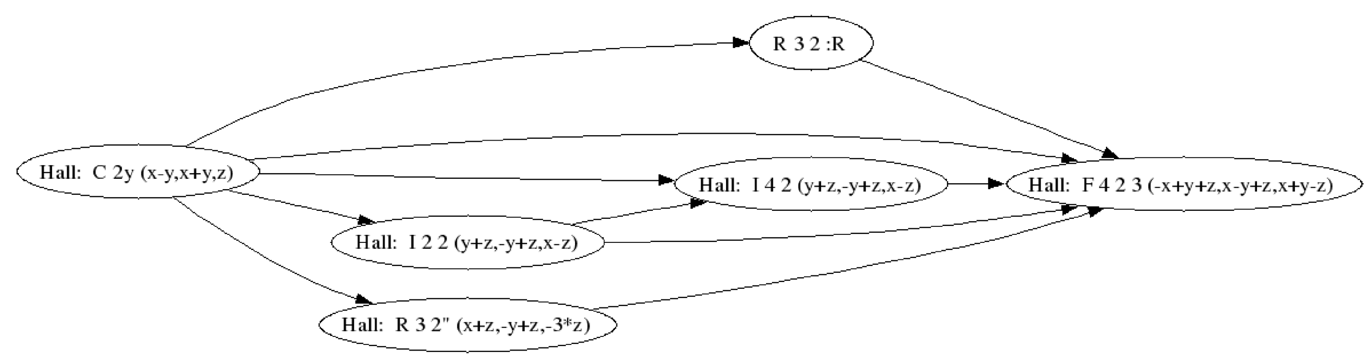

Figure 3. A point group graph given a user supplied unit cell and space group.

7. Acknowledgments and availability

We gratefully acknowledge the financial support of PHENIX industrial consortium and the NIH/NIGMS through grants 5P01GM063210 and 5P50GM062412. Our work was supported in part by the US department of Energy under Contract No. DE-AC02-05CH11231.

Xtriage is part of the PHENIX suite for macromolecular crystallography and is available at `http://www.phenix-online.org`. The source code of *Xtriage* is part of the open source crystallographic libraries *CCTBX*, available at `http://cctbx.sf.net`.